Published in: Mayr, Philipp/Tudhope, Douglas/Golub, Koraljka/
Wartena, Christian/William De Luca, Ernesto (eds.): Proceedings of the 15th European
Networked Knowledge Organization Systems Workshop (NKOS 2016). CEUR
Workshop Proceedings, Vol. 1676. Pp. 59-63.

Re-designing Online Terminology Resources for German Grammar

Project Report

Karolina Suchowolec, Christian Lang, and Roman Schneider

Institut für Deutsche Sprache (IDS), Mannheim, Germany
{suchowolec, lang, schneider}@ids-mannheim.de

Abstract The compilation of terminological vocabularies plays a central role in the organization and retrieval of scientific texts. Both simple keyword lists as well as sophisticated modellings of relationships between terminological concepts can make a most valuable contribution to the analysis, classification, and finding of appropriate digital documents, either on the Web or within local repositories. This seems especially true for long-established scientific fields with various theoretical and historical branches, such as linguistics, where the use of terminology within documents from different origins is sometimes far from being consistent. In this short paper, we report on the early stages of a project that aims at the re-design of an existing domain-specific KOS for grammatical content *grammis*. In particular, we deal with the terminological part of *grammis* and present the state-of-the-art of this online resource as well as the key re-design principles. Further, we propose questions regarding ramifications of the Linked Open Data and Semantic Web approaches for our re-design decisions.

1 State of the Art

Grammis is a specialized hypertext resource hosted by the Institute for German Language (IDS) in Mannheim.[1] It brings together terminological, lexicographical, and bibliographic information about German grammar. Initiated more than two decades ago, it combines the traditional description of grammatical structures with the results of corpus-based studies. From a technical point of view, all primary data and meta data is coded within more than one thousand semi-structured XML instances that are composed of semantical markup element types (*title*, *header*, *example*, *link anchor*, etc.).

From the user's perspective, *grammis* consists of several components. The core component uses comprehensive reference texts, which describe grammatical phenomena in great detail. In contrast, the terminological component is meant to be a short reference, where each entry concisely describes a grammar concept and, therefore, gives just the basic notion of it. It also (statically) points the user to relevant entries in the core component for further reference.

[1] http://hypermedia.ids-mannheim.de/

As of now, the intended user of *grammis* needs a sound knowledge in the field of grammar. This means that he or she needs to be a professional linguist, most likely a grammar scholar. However, there are other online resources on German grammar hosted by IDS that target at different user groups; for instance, *ProGr@mm* is dedicated to teaching and explaining grammar to non-scholars, such as students of linguistics.[2] It mirrors the structure of *grammis*, i.e. has similar components such as comprehensive references texts and terminological reference. These components, however, may or may not have the same content as corresponding *grammis* entries. Also, there are other, more specialized, resources, for example on so-called *connectors*, which are results of different research projects at IDS.

Finally, there is a separate terminological resource called *grammatische Ontologie* [9].[3] It is, in fact, a taxonomy of regular hierarchical relations such as broader/narrower term (both generic and partitive) (BT/NT), related term (RT) as well as synonym relation. No other information on terms and concepts such as scope notes or definitions is given. This resource, which has been developed and maintained independently of *grammis*, is implemented in an object-relational database management system (ORDBMS). It also serves a different purpose—it (dynamically) generates a list of references to other terminological, lexicographical, and bibliographic resources by IDS on a given topic, and uses the hierarchy to generate more relevant hits.

To sum it up, the landscape of online terminology resources at IDS is heterogeneous. Above all, it is a result of different research projects, with different goals, scopes, scholarly traditions, and persons involved. Therefore, it covers different areas of grammar with different degrees of specialization. Moreover, terminology is currently managed within different tools, depending on whether it is used in the *grammis* dictionaries, ontology, or bibliography. Further, the content often reflects needs of heterogeneous user groups. And finally, different resources were designed with different purposes in mind—to serve as a concise grammar reference or a repository for enhancing information retrieval.

In addition, there is a broad spectrum of terminology that is yet to be covered by terminology resources. To deal with these heterogeneities, distributions, and unsatisfactory coverage of terminology resources, IDS has launched a project for re-designing the current terminology management.

2 Re-Design Principles

To address the above-mentioned issues of heterogeneity, distribution, and coverage of the current terminology resources, we propose the following work packages for the re-design of the terminology management, which are described in more detail below.

– Combining distributed terminology resources into one resource;

[2] http://hypermedia.ids-mannheim.de/programm/
[3] http://hypermedia.ids-mannheim.de/call/public/termwb.html

– Updating the content using automated keyword extraction techniques;
– Ensuring interoperability of the new resource with other projects;
– Implementing a new backend for terminology management.

2.1 Combining Resources into one Resource

One way of improving the current terminology management for both, the user and the (terminology) author, is to unite the scattered terminology resources into one global resource. The new resource should first incorporate not only the descriptions of single concepts and terms, but also the relations between them. In other words, it should contain both—the hierarchy and the descriptive information about terms and concepts within this hierarchy.

Moreover, we are exploring the question, to what extent we can further unify this resource to globally serve for different target groups and for results from different research projects. These considerations have implications for the content of the entries, but also for their structure and, derived from that, for the data modelling. Therefore, we are currently evaluating the style and the structure of the existing terminology entries in order to find a common and hence more consistent way of authoring them.

2.2 Automated Keyword Extraction

To enhance the coverage of the new resource, we use automatic keyword extraction on the core component's entries. Topic Rank [1], an algorithm based on Page Rank [2], is applied to each entry of *grammis*, *ProGr@mm* and the specialized resources. In addition to standard linguistic preprocessing we optimize the algorithm's input by excluding non-domain-specific (loglikelihood [7] and weirdness ratio [3] against DeReKo corpus [cf. 6]) and non-characteristic candidate words (TF-IDF [10], Gries DP [4]). As a plus, we exploit the semantic markup of heterogeneous text sections, coded with XML element types. In a preliminary study, we compare recall and precision performance of the automatized extraction against a human annotated gold standard.

2.3 Interoperability

The end-user of the new resource will be provided, as with the current resource, with an online interface. However, we also want to ensure that our data is transparent as well as easily accessible, exchangeable, and reusable within different (scientific) contexts and applications. In particular, we want to make it available to the communities that provide the scientific backbone of our project i.e. the terminology community and the taxonomy and thesaurus community. Therefore, we implement standard exchange formats for our data.

For the terminology community, such standard is TBX [5]. As of now, we are evaluating data categories available in TBX and mapping them to our current data categories. Implementing TBX without information loss would possibly

mean flattening our current meta model [cf. 8], hence reducing the number of available relation types to fit into the three-level TBX meta model.

For the taxonomy community, we identified SKOS as a possible format [11]. We also would like to look into Lemon as an alternative format with more linguistic power [12]. We are exploring the implications of these formats on our meta model and section 3 deals with our first considerations on SKOS.

2.4 New Backend for Terminology Management

From all of the above follows that the current terminology backend tools need to be reconsidered in order to account for the changes in the terminology management. After specifying the requirements, we are now evaluating different options for the new tool, looking into commercial native terminology, native thesaurus, and hybrid solutions, but also considering a re-design of our own tools. Most importantly, the new tool should manage both the hierarchy and the descriptive information on concepts and terms within this hierarchy. Since no language management for grammatical terminology is intended, it also needs to efficiently manage quasi-synonyms, i.e. partially-equivalent terms, accounting for different schools of linguistics. Further features comprise the support of the above-mentioned exchange formats and the interoperability with other in-house applications. Finally, visualizing data as a graph is a desirable, but an optional feature.

3 Sematic Web and Linked Open Data Ramifications

As of now, we are still in the early stages of our project and there are some questions we would like to discuss. As mentioned above, we are assessing the implications of SKOS for our project, which can be summarized by the following questions:

- What are the challenges when converting data into SKOS?
- What implications does SKOS have for our meta model?
- Taking it further, what implications for our meta model would becoming a part of Semantic Web and Linked Open Data have?

References

1. Bougouin, A., Boudin, F., Daille, B.: TopicRank: Graph-based topic ranking for keyphrase extraction. In: International Joint Conference on Natural Language Processing. pp. 543–551 (2013)
2. Brin, S., Page, L.: The anatomy of a large-scale hypertextual search engine. Computer Networks and ISDN Systems 30(1-7), 107–117 (1998)
3. Gillam, L., Tariq, M., Ahmad, K.: Terminology and the construction of ontology. Terminology 11(1), 55–81 (2005)
4. Gries, S.T.: Dispersions and adjusted frequencies in corpora. International Journal of Corpus Linguistics 13(4), 403–437 (2008)

5. ISO 30042: Systems to manage terminology, knowledge and content – TermBase eXchange TBX. First edition (2008)
6. Kupietz, M., Keibel, H.: The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research. In: Minegishi, M., Kawaguchi, Y. (eds.) Working Papers in Corpus-based Linguistics and Language Education, pp. 53–59. No. 3, Tokyo University of Foreign Studies, Tokyo (2009)
7. Rayson, P., Garside, R.: Comparing corpora using frequency profiling. In: Proceedings of the workshop on Comparing Corpora, 38th annual meeting of the Association for Computational Linguistics (ACL). pp. 1–6. Hong Kong (2000)
8. Schneider, R., Gottron, T.: A hybrid approach to statistical and semantical analysis of Web documents. In: Proceedings of the IASTED International Conference Internet and Multimedia Systems and Applications (EuroImsa). pp. 115–120 (2009)
9. Sejane, I.: Wissensrepräsentation Linguistik. Modellierung, Potenzial und Grenzen am Beispiel der Ontologie zur deutschen Grammatik im GRAMMIS-Informationssystem des IDS, Mannheim. phdthesis, Ruprecht-Karls-Universität Heidelberg (2010)
10. Spärck Jones, K.: A statistical interpretation of term specifity and its application in retrieval. Journal of Documentation 28(1), 11–21 (1972)
11. W3C: SKOS Simple Knowledge Organization System reference (2009), `https://www.w3.org/TR/2009/REC-skos-reference-20090818/`, 2009-08-18
12. W3C Ontology Lexicon Community Group: Final model specification (2016), `https://www.w3.org/community/ontolex/wiki/Final_Model_Specification`